Regularized Extended Skew-Normal Regression

K. Shutes & C.J. Adcock

September 9, 2014

Abstract

This paper considers the impact of using the regularisation techniques for the analysis of the (extended) skew normal distribution. The models are estimated using Maximum Likelihood and compared to OLS based LASSO and ridge regressions in addition to non- constrained skew normal regression. The LASSO is seen to shrink the model's coefficients away from the unconstrained estimates and thus select variables in a non- Gaussian environment.

1 Introduction & Motivation

Variable selection is an important issue for many fields. It is also noticeable that not all data conforms to the standard of normality. This paper addresses the issue raised by Bühlmann [2013] of the lack of non-Gaussian distributions using regularisation methods. Within the statistics literature there are many applications of penalised regressions. There are other fields such as finance and econometrics where these approaches are less common. This paper extends this to consider situations where shrinkage of the coefficients might be helpful and one has an a priori expectation of non-normality in the data.

Variable selection is an important part of the modelling process. A number of approaches such as stepwise regression or subset regression have previously been used with metrics such as Aikake Information Criteria (Akaike [1974]) used as the decision criterion. There are well documented problems with these approaches. The use of regularised regressions mitigate these problems. The coefficients are shrunk towards zero, which creates a selection process.

In the majority of cases, the use of the regularisation techniques are based upon Gaussian distributed errors and Ordinary Least Squares. Though in many cases this is sufficient, there are many cases such as those in finance where normality is not an appropriate assumption. This paper looks to add to the regularisation literature by extending the Least Absolute Shrinkage & Selection Operator (henceforth LASSO (Tibshirani [1996]) to accommodate shrinkage within the higher moments via the use of the extended skew normal based regression model (Adcock & Shutes [2001] & Shutes [2004]). The method proposed here uses the technique of the LASSO, i.e. the introduction of ℓ_1 norms, but in contrast to the literature based on Gaussian regression, further norms

are introduced on the skewness parameters. This will imply that in addition to the variable selection made via the standard approach the method also performs a selection of non-normality as the extra parameters control the skewness and kurtosis.

The rest of the paper is organised as follows. A consideration of the extended skew normal and the LASSO is presented with the relevant estimation and an example to conclude. A standard data set from the machine learning literature, that of diabetes patients is used (see Efron et al. [2004] where it is more fully described). All estimation was performed in R [2008] with package Azzalini [2013].

2 Literature Review & Definitions

2.1 Regularization

In many fields, regularisation has a substantial history. In circumstances of ill-formed problems, such as multi-collinearity or non-full rank in the independent variable matrix, it is possible to use these approaches. Ridge regression is perhaps the best known example (for example Hoerl & Kennard [1970]), where the problem of multicollinearity is dealt with by the imposition of a constraint on the coefficients of the regressions. This estimator is known to be biased however it is the case that the approach gives estimators with lower standard errors. The ridge and the LASSO exhibit an equivalence between the penalty formulation and that of a Lagrangean, with a correspondence between the Lagrange multiplier and the value of ϵ as shown in Osborne et al. [1999]. The penalised function for the estimation is given by:

$$\begin{aligned}
\beta_R &= \arg\min_{\beta} \left(Y_i - \beta_0 - X_i\beta^T\right)^T \left(Y_i - \beta_0 - X_i\beta^T\right) \quad \text{s.t.} \quad \beta^T\beta \leq \epsilon \qquad (1)\\
&= \arg\min_{\beta} \left(Y_i - \beta_0 - X_i\beta^T\right)^T \left(Y_i - \beta_0 - X_i\beta^T\right) + \nu\beta^T\beta \\
&= \left(X^TX + \nu I\right)^{-1} X^Ty
\end{aligned}$$

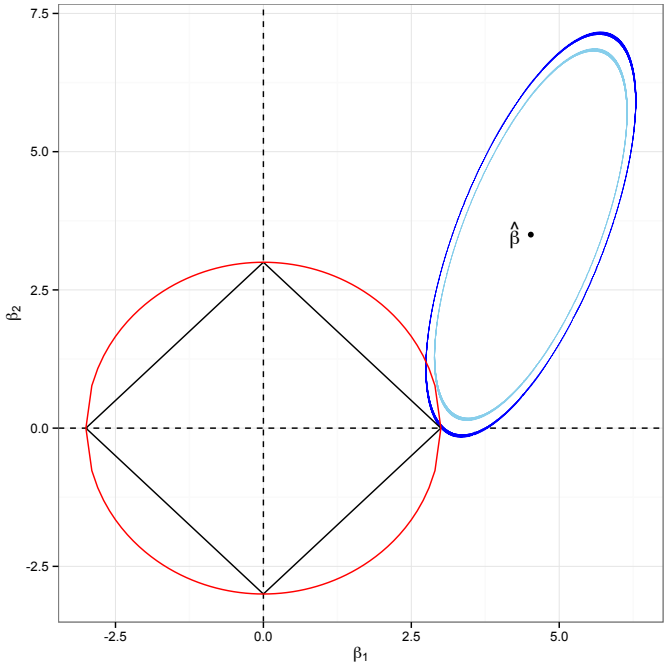
This approach does not perform any form of variable selection as, although it does shrink coefficients, it does not shrink them to 0. The ν parameter[1] acts as the shrinkage control with $\nu = 0$ being no shrinkage and therefore ordinary least squares. This can be compared to the Least Absolute Shrinkage & Selection Operator (LASSO). In this case the penalty is based on the ℓ_1 norm rather than the ℓ_2 norm of the ridge approach. Hence the problem becomes:

$$\begin{aligned}
\beta_L &= \arg\min_{\beta} \left(Y_i - \beta_0 - X_i\beta^T\right)^T \left(Y_i - \beta_0 - X_i\beta^T\right) \quad \text{s.t.} \quad \mid \beta \mid \leq \epsilon \qquad (2)\\
&= \arg\min_{\beta} \left(Y_i - \beta_0 - X_i^T\beta\right)^T \left(Y_i - \beta_0 - X_i^T\beta\right) + \nu \mid \beta^T \mid 1
\end{aligned}$$

[1] Traditionally the Lagrangean multiplier is denoted λ, however due to the use of λ as the skewness parameter in the distribution, the Lagrangean is denoted ν throughout this paper.

The variable selection property is clearly shown graphically in Figure 1 when considering two parameter estimates, with the LASSO (black) and ridge (red). The estimator loss functions are shown as ellipses. The point of tangency are the estimates for each

Figure 1: Differences Between LASSO & Ridge Regressions



technique. The LASSO shrinks β_1 to 0, whereas the ridge regression approaches it. The OLS estimator is given as $\hat{\beta}$. The parameter ν controls the amount of penalty applied to the parameters for the LASSO. Fu and Knight [2000] show that under certain regularity conditions, the estimates of the coefficients are consistent & that these will have the same limiting distribution as the OLS estimates.

There is a generalisation such that the γ-th norm is used. This is the bridge estimator. The γ-th norm is defined as:

$$|| \beta ||_{\gamma} = \left(\sum | \beta_i |^{\gamma} \right)^{\frac{1}{\gamma}} \tag{3}$$

This therefore implies that the bridge regression, despite first impressions will not select variables unless $\gamma < 1$ in which case the penalty function is non-concave and the estimates may not be unique, though they may be set at zero. These estimators, LASSO, bridge and ridge are all forms of Bayesian estimator with priors based on a LaPlace or variants of this based on a log exponential function.

2.2 The Skew Normal Distribution

The skew skew normal distribution has become increasingly well used within a number of fields since its initial description by Azzalini [1985]. A particularly attractive feature of the distribution is that it includes the Gaussian as a limiting case. In its simplest form the distribution is described by the following density function:

$$\begin{aligned} h(y) &= 2\phi(y)\,\Phi(\lambda y) \\ -\infty < &\ \lambda\ < \infty \\ -\infty < &\ y\ < \infty \end{aligned} \tag{4}$$

with λ controlling the degree of skewness of the distribution. The case λ=0 will lead to a standard normal distribution.

Azzalini [1985] & [1986] proposes that the skew normal distribution is best thought of as a combination of a symmetric element and a skewing element, which is a truncated normal distribution with mean of 0. This is generalised in Arnold & Beaver [2000] and Adcock & Shutes [2001] where the truncated normal has a mean of τ. Thus the density function can be written as:

$$f(r) = \frac{1}{\Phi(\tau)}\phi\left(r; \mu + \lambda\tau, \sigma^2 + \lambda^2\right)\Phi\left(\frac{\tau + \frac{\lambda}{\sigma^2}(r-\mu)}{\sqrt{1+\frac{\lambda^2}{\sigma^2}}}\right) \tag{5}$$

where ϕ and Φ are the probability density and cumulative functions of the normal distribution respectively.

It is possible to use the following parameterization, with γ and ω^2 being the mean and the variance of the normal part of the distribution respectively:

$$\begin{aligned} \gamma &= \mu + \lambda\tau \\ \omega^2 &= \sigma^2 + \lambda^2 \\ \psi &= \sqrt{\sigma^2+\lambda^2}\frac{\lambda}{\sigma} = \omega\frac{\lambda}{\sigma} \\ \frac{\psi^2}{\omega^2} &= \frac{\lambda^2}{\sigma^2} \end{aligned} \tag{6}$$

This parameterisation allows a simpler description of the distribution. This is not a unique transformation. However the definitions used are easily extendable to the multivariate distribution. The probability density function can be expressed in terms of these parameters as:

$$f_R(r) = \frac{1}{\Phi(\tau)}\phi(r;\gamma,\omega^2)\Phi\left(\tau\sqrt{1+\frac{\psi^2}{\omega^2}} + \frac{\psi}{\omega^2}(r-\gamma)\right) \tag{7}$$

where $\phi(x;\mu,\sigma^2)$ is the probability density function of a normally distributed variable with mean μ and variance σ^2. This gives an extension to the standard skew normal distribution, known as the extended skew normal.

The application of the LASSO type approach to the skewed family of distributions is limited. Wu et al. [2012] consider the variable selection problem for the skew normal family. However they use a fixed but estimated skewness parameter in essence removing the skewness problem in conjunction with a quadratic expansion of the penalised likelihood to give a tractable solution. Their focus is very much on the location and scale parameters rather than the skewness with a view to modelling the variance as an entity as well as the mean i.e. regression style models. The penalised likelihood approach used both in Wu and here is found in Fan and Li [2001]. This allows both the estimation and standard errors to be estimated despite the singularity introduced by the constraint.

3 Likelihood Functions

In order to use the LASSO style estimators, it is necessary to consider the relevant likelihood estimators in light of the constraints. We can think of the constrained likelihood as having two elements, the objective and the constraint.

The likelihood function of the extended skew normal distribution is somewhat non-linear. Using the specification above, the likelihood is given by:

$$\begin{aligned}\ell_i(y;\tau,\gamma,\beta,\psi,\omega^2) &= -\ln\Phi(\tau)-\frac{1}{2}\ln\omega^2-\frac{1}{2}\ln 2\pi-\frac{1}{2\omega^2}\left(y_i-\beta_0-\beta x_i-\gamma\right)^2 \\ &\quad +\ln\Phi\left(\tau\sqrt{1+\frac{\psi^2}{\omega^2}}+\frac{\psi}{\omega^2}(y_i-\beta_0-\beta x_i-\gamma)\right)-\nu_1\left(||\ \beta\ ||_1+||\ \psi\ ||_1+||\ \tau\ ||_1\right)\end{aligned} \tag{8}$$

This is the standard log-likelihood function for the extended skew normal with the addition of the LASSO penally for the coefficients and the skewness parameter.

The regression coefficients where the constraints can potentially bind are given below.

$$\frac{\partial\ell}{\partial\beta} = \frac{x}{\omega^2}(y-\beta x-\gamma)-\frac{\psi}{\omega^2}x\zeta_1\left(\tau\sqrt{1+\frac{\psi^2}{\omega^2}}+\frac{\psi}{\omega^2}(y-\beta x)\right)-\mathrm{sgn}(\beta)\nu_1 \tag{9}$$

4 Estimation

For Gaussian based estimations it is possible to leverage the co-ordinate descent approach to update the estimates of the relevant coefficients until convergence to the LASSO solution occurs. Assuming uncorrelated predictors, the updating procedure can be based on the product of the residuals and the relevant predictors and the value of the Lagrange multiplier. This produces a whole path solution with the different solutions for the problem providing the starting point for the next optimisation thus reducing the issues with convergence[2] and speed. The approach taken here is to use direct estimation of the likelihood function for the distributions where τ is unconstrained (the extended skew normal) and where it is constrained to $\tau = 0$, the skew normal. Each estimator used

[2]As noted in Azzalini and Capitanio [1999] the likelihood function of the skew normal is not convex in its standard form.

the previous estimate as the starting point of the algorithm to increase the speed of the estimation.

4.1 Estimation with Maximum Likelihood

Estimation was performed using a maximum likelihood approach with the nuisance parameter, ν being based on a grid in the first case and then cross validation being used to optimise the choice of this parameter. Using the non-constrained maximum likelihood estimates as the initial points to aid in convergence, the estimations were performed with a transformation of the parameter ν to $\exp(\nu)$. This leads to more satisfactory convergence of the algorithms and allowed a greater range of the parameter than a simple linear constraint would allow.

The estimation of ν used a 10-fold cross-validation over an identical grid of ν parameter values. The CV errors are calculated off the hold-out sample of this, with the ν selected by the min+1S.E. rule of thumb being used as a fixed parameter within the final, whole sample estimation. Thus the process involves sampling in order to estimate the nuisance parameter, with that value then being used to select the model using the whole data set.

5 Data & Maximum Likelihood Estimation

The data used was a standard machine learning example, the diabetes dataset (from Efron et al. [2004]). These relate the progress of diabetes over a year to the age, weight, BMI and various serum measurements. There are 442 observations with the first non-interaction terms used. The data are standardised to have 0 mean and an unit ℓ_2-norm. Though this is not a $p >> n$ situation, it serves to demonstrate the technique and places this approach in the corpus of penalised regression.

This is supplemented with a set of simulations based on 10 variables with lengths of 10000 and 1000 observations. Fifty different sets of data are used to demonstrate the properties of the estimation. This is performed with the mean and standard error of the mean, median and 25th and 75th quantiles given.

5.1 Simulated Data

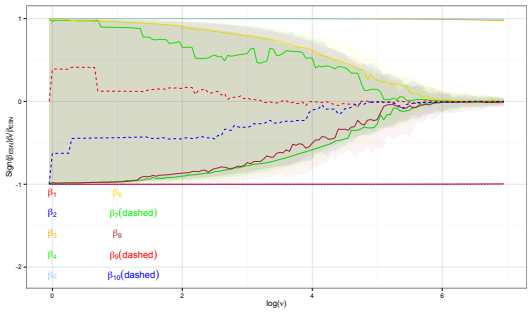
Fifty simulated data sets of both 1000 and 10000 observations were created with specific seeding points. These contained 10 independent variables. The data generating process was identical for all of the simulations ($\beta_1 < 0$, $\beta_2 < 0$, $\beta_3 > 0$ and $\beta_5 > 0$ are all non-zero). The estimates are reported as a proportion of the full Maximum Likelihood estimators.

Those variables that are included in the data generating process are stable around the MLE coefficients (qv. Figures 2a and 4a), whereas those omitted from the data generating process are restricted and converge to zero (Figure 2a & 4a) and have a mean of zero (Figures 2b & 4b). These have a wider dispersion than the variables included in the data generating process.

Results for both the lengths are similar in substance, though the dispersion is higher in the smaller data sets. In both cases the skewness parameters (γ and τ) converge to zero as the penalty increases even though the actual value is not zero (Figures 3a and 5a with mean values shown in Figures 3b and 5b). This is in part due to the non-linearities associated with the likelihood function. The instability that this creates gives a median value of zero. The model is penalising the asymmetry and removing it from the regression in these cases.

Figure 2: Paths of LASSO Coefficients for the Skew Family of Distributions for the Simulated Data

(a) LASSO Regression Coefficients (β) of Variables by ν (N=10000)



(b) Mean LASSO Regression Coefficients (β) of Variables by ν (N=10000)

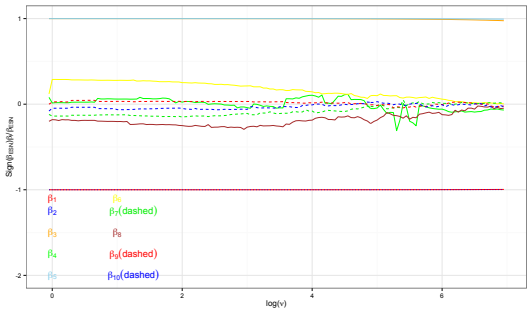
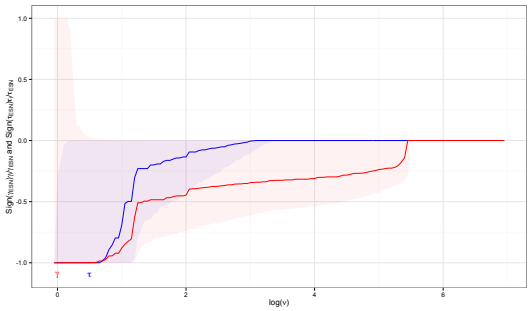
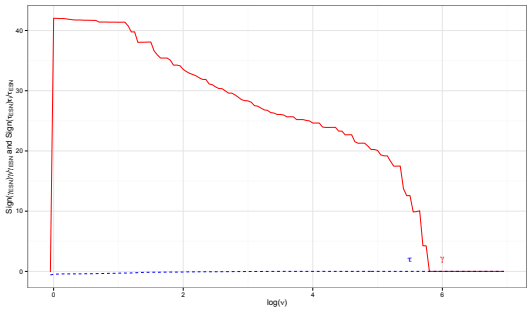


Figure 3: Skewness Parameter Estimates

(a) Skewness Parameter Estimates of Simulation Data (N=10000)



(b) Mean Skewness Parameter Estimates of Simulation Data (N=10000)

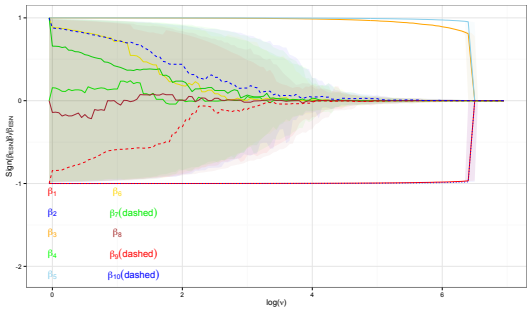


5.2 Diabetes Data

The results are presented with skew normal ($\tau = 0$) and extended skew normal ($\tau \leqslant 0$), Gaussian LASSO and Ridge regressions in Table 1. The Maximum Likelihood approach used a grid of Lagrange multipliers and the coefficients from each of these values are

Figure 4: Paths of LASSO Coefficients for the Skew Family of Distributions for the Simulated Data

(a) LASSO Regression Coefficients (β) of Variables by ν (N=1000)



(b) Mean LASSO Regression Coefficients (β) of Variables by ν (N=1000)

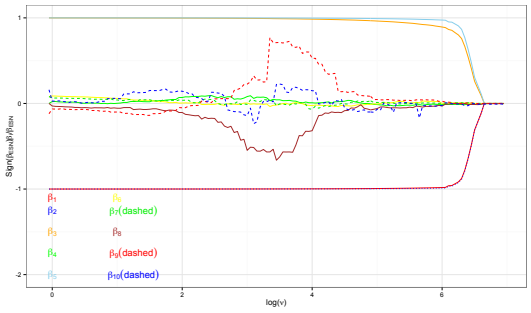
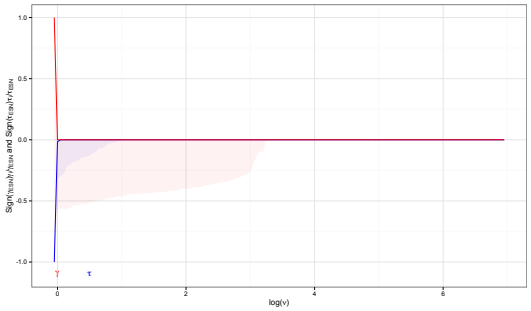
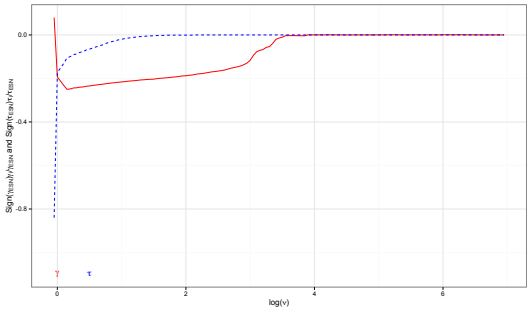


Figure 5: Skewness Parameter Estimates

(a) Skewness Parameter Estimates of Simulation Data (N=1000)



(b) Mean Skewness Parameter Estimates of Simulation Data (N=1000)



recorded. These are presented graphically in Figures 6a and 6b with the coefficients presented as a proportion of the unconstrained maximum likelihood estimates[3]. As can be seen the estimates converge to zero as the penalty increases. A number of coefficients were somewhat unstable for the skew normal, though this is less problematic for the extended skew normal estimations. This is due to the relative smoothness of the likelihood functions under specific conditions (examples are given in Azzalini and Capitanio [1999]).

The path of the regression coefficients are given in Figure 6a and 6b using a grid-based path. These are given as a proportion of the unconstrained estimates (with a sign modification to aid visualisation). These diagrams show the variable selection ability of the LASSOs.

The LASSO parameter, ν is selected using the 10-fold cross validation. Using the rule of thumb that one should maximise the cross validated parameter within a standard error (Breiman et al. [1984]) of the MSE of the minimum, the optimal value of $\ln(\nu)$ is -3.4 for the skew normal and -3.6 for the extended version as is shown in Figure 7a & 7b respectively. The relevant ν parameters are shown in Figures 6a & 6b as the vertical dashed line. These results demonstrate that there is variable selection under both the skew normal and the extended skew normal LASSOs. The regression coefficients have a similar path for each of the distributions, though not identical.

The selection implies that the variables 2, 3, 4, 7 and 9 are to be included in the skew normal model model with the other coefficients being less than 1% of their standard MLE estimate with variable 10 also included in the ESN LASSO as in the case of the Gaussian LASSO. The skew normal LASSOs do not include variable 5 unlike the standard LASSO.

The parameters associated with the skewness, λ and τ, are estimated from the likelihood function. These are presented below in Figure 8a & 8b. What is immediately obvious is that the skewness parameter under the non-extended formulation is erratic, whereas under the extended form there is more direct convergence.

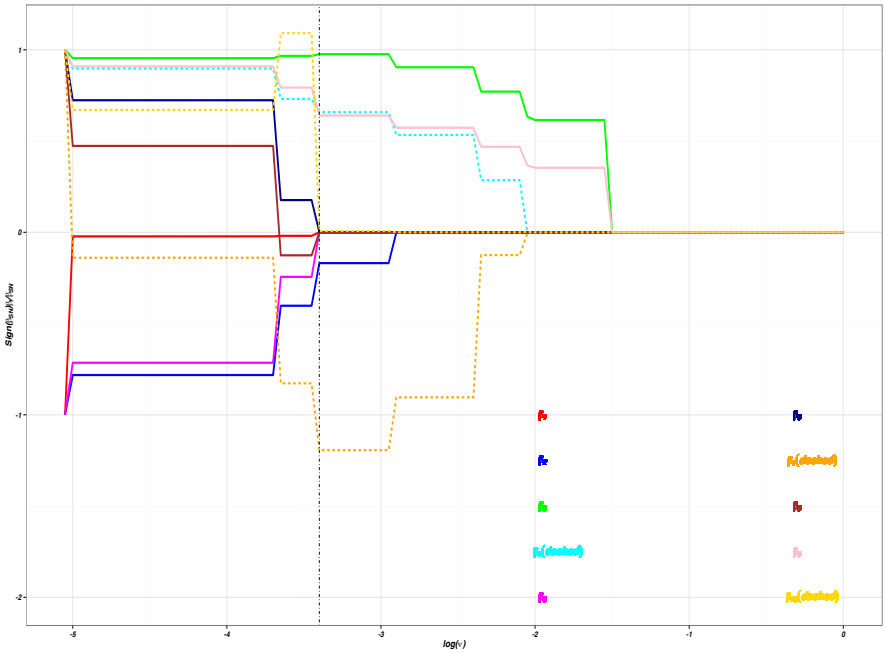
The underlying distributions with λ & τ at the cross validated parameter value of ν is shown in Figure 9. As can be seen, the distributions are very similar. The ratio of the extended to unextended variants has a range of 5% and has a maximum difference in the tails. This is to be expected as the extended skew normal's extra parameter does allow control over the tails of the distribution.

The OLS ridge regression shrinks the coefficients towards 0 however this is not as extreme as that of the LASSO in both the Gaussian and non- Gaussian scenarios. The (leave one out) cross validated LASSO Gaussian coefficients are also given in Table 1. These were estimated using `glmnet` (Friedman et al. [2010]). The penalty for the ridge regression is selected using the approach of Cule and De Iorio [2012] based on cross-validation. There is more shrinkage under the skew normal approaches to the LASSO. Thus the skew normal creates a more parsimonious regression but the skewness parameters are non-zero. There is therefore a trade-off between a more parsimonious

[3]Given that the LASSO parameter is re-parameterized as $\exp^{\nu}$, the unconstrained optimum is given as a small step away from the start of the grid search in order to demonstrate the shrinkage across the range.

Figure 6: Path of LASSO Coefficients for the Skew Family of Distributions

(a) Path of Skew Normal LASSO Regression Coefficients (β) by ν



(b) Path of Extended Skew Normal LASSO Regression Coefficients (β) by ν

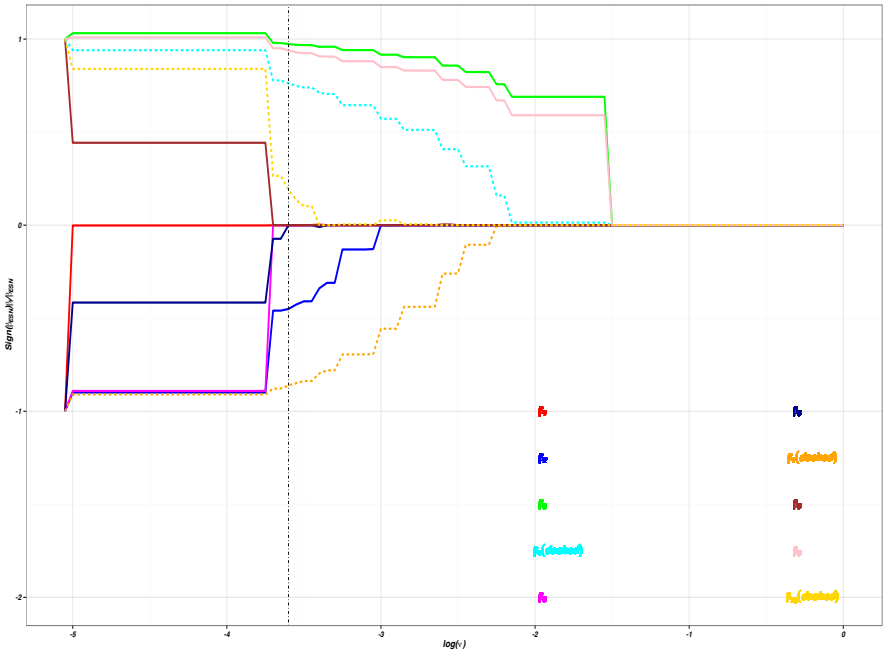
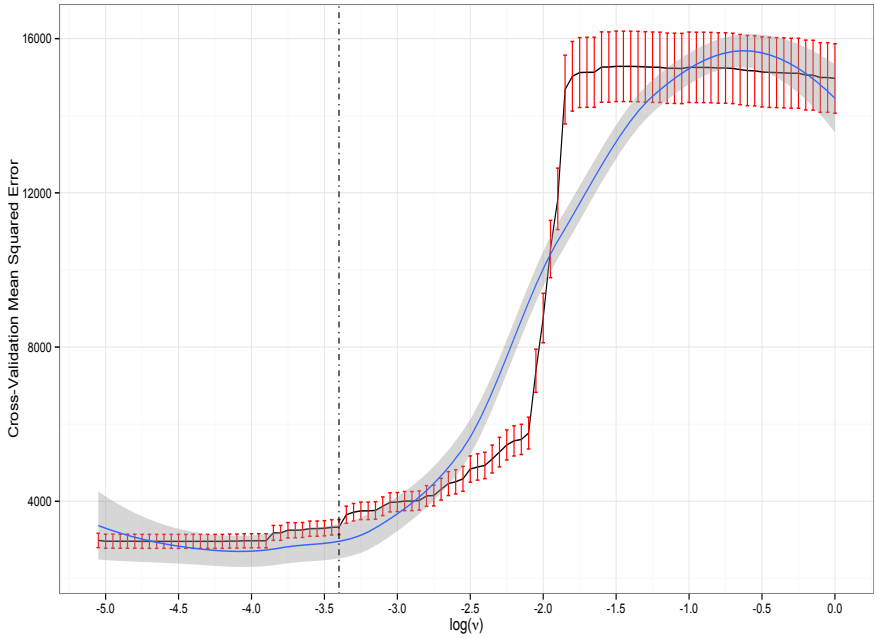
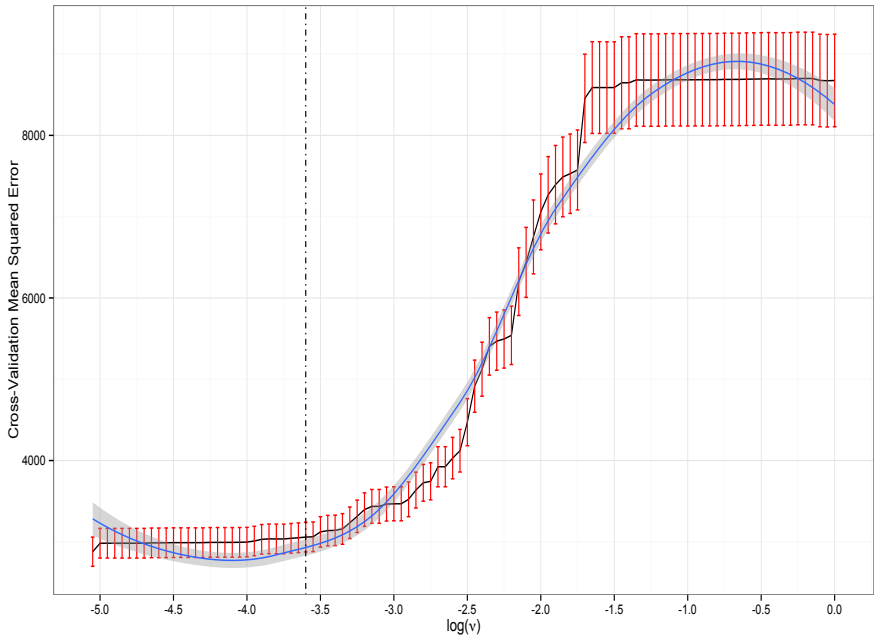


Figure 7: Cross Validation Results for the Selection of ν, the LASSO parameter for Skew and Extended Skew Normal

(a) Cross Validation of Skew Normal LASSO



(b) Cross Validation of Extended Skew Normal LASSO

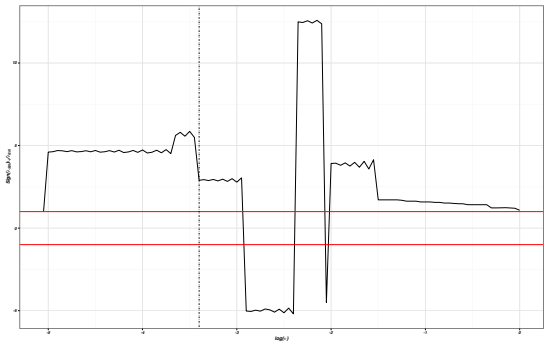


regression and a parsimonious distribution.

The skew parameters are acting to counter-act the variable not included.

Figure 8: Skewness Parameters for the Skew and Extended Skew Normal

(a) Path of Skewness Parameter λ for the Skew Normal LASSO



(b) Path of Skewness Paramters λ & τ for the Extended Skew Normal LASSO

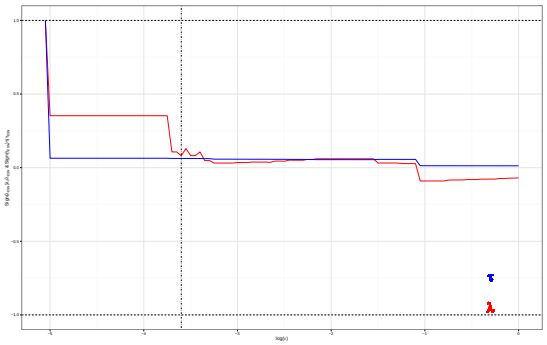
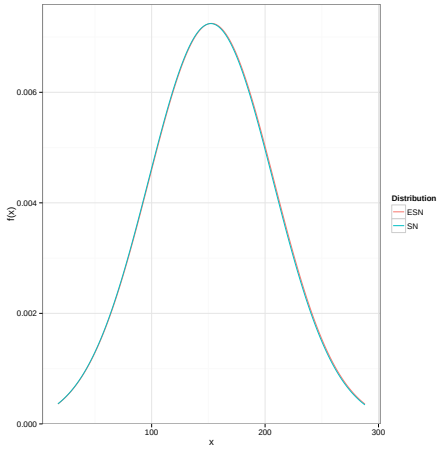


Figure 9: Distribution of Skew Normal Distributions from Parameter Estimates

(a) Distribution of the Skew Normal & Extended Skew Normal at estimate of ν



(b) Ratio of Distribution of Skew Normal & Extended Skew Normal Distributions

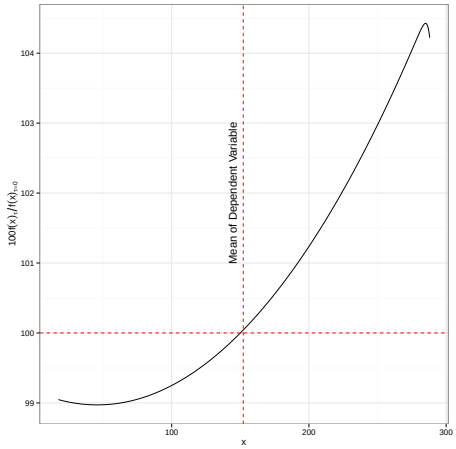


Table 1: Estimates of the Skew Normal LASSO for Diabetes Data

	SN LASSO	ESN LASSO	SN MLE		ESN MLE		LASSO	Ridge		OLS	
	Coef	Coef	SN	SE	ESN	SE	CV.LASSO	Ridge	Ridge SE	OLS	OLS SE
μ	151.487	152.719	152.1335	2.544	152.133	2.552	152.133	152.133	NA	152.133	2.576
β_1	-	-	-10.012	59.297	-7.231	59.191	-	-4.816	57.599	-10.012	59.749
β_2	-40.484	-105.654	-239.819	61.070	-234.654	60.651	-196.053	-228.124	58.710	-239.819	61.222
β_3	507.480	514.916	519.840	65.816	529.039	65.915	522.070	515.391	63.156	519.840	66.534
β_4	213.516	244.548	324.390	64.804	320.971	64.811	296.268	316.125	62.340	324.390	65.422
β_5	-	-	-792.184	414.036	-101.146	415.433	-102.047	-206.171	102.045	-792.184	416.684
β_6	-	-	476.746	337.776	-84.037	338.006	-	13.835	99.620	476.746	339.035
β_7	-120.622	-170.463	101.045	209.892	-197.840	211.517	-223.27	-150.203	91.810	101.045	212.533
β_8	-	-	177.064	159.876	106.878	160.037	-	115.787	114.508	177.064	161.476
β_9	480.669	458.722	751.279	170.958	488.531	171.222	513.684	518.312	76.632	751.279	171.902
β_{10}	-	13.586	67.625	65.334	70.653	65.368	53.937	75.172	63.061	67.625	65.984
ν	0.033	0.027									
$\log(\nu)$	-3.4	-3.6									
λ	0.014	-9.627	0.005	0.101	-118.631	0.000					
σ	55.081	55.237	53.476	1.799	53.648	1.816					
τ	0	2.710	43.242	0.000							
lp	-2444.44	-2434.91	-2385.99		-2387.43						

Key:
ESN LASSO= Estimation of Extended Skew Normal LASSO with coefficients greater than 1% of ESN MLE
SN LASSO= Estimation of Skew Normal LASSO with coefficients greater than 1% of SN MLE
SN MLE= Estimation of Skew Normal by MLE
LASSO= Gaussian based LASSO with penalty parameter estimated using Cross Validation
Ridge= Gaussian based Ridge with penalty parameter estimated using Cross Validation
OLS= Gaussian based regression

6 Conclusions

The skew normal is an example of a well developed class of asymmetric distributions. This paper has shown that it is possible to adapt the estimation of regressions based on this distribution to include a LASSO type penalty. This is seen to shrink the estimates of regression coefficients and thus perform a variable selection role. This therefore allows the analysis of data using a non- Gaussian toolbox and thus address the issue raised by Bühlmann [2013]. Natural extensions from this work include a generalisation from the skew normal distribution to include other, spherically symmetric distributions. These, such as the skew Student distribution would increase the application of these approaches to situations where higher moments are critical such as finance. Further the extension of the LASSO to its generalisation of the elastic net is also possible as is the Bayesian estimation using double exponential priors on the regularised coefficients.

The skew normal family of LASSOs will trade off the distribution complexity with the regression complexity relative to the Gaussian distribution. The skewness parameters act in the same manner fundamentally as the regression coefficients with the approach constraining them towards 0 as the penalty increases. Thus the Gaussian and the skewed variants will converge if the skewness parameters are driven towards 0 relatively soon in the process.

References

C. J. Adcock and K. Shutes. Portfolio Selection Based on The Multivariate Skew-Normal Distirbution. In A Skulimowski, editor, *Financial Modelling*. Progress and Business Publishers, 2001.

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716 – 723, dec 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.

B. C. Arnold and R. J. Beaver. Hidden Truncation Models. *Sankhya, Series A*, 62 (22-35), 2000.

A. Azzalini. A Class of Distributions which Includes The Normal Ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.

A. Azzalini. Further Results on a Class of Distributions which Includes The Normal Ones. *Statistica*, 46(2):199–208, 1986.

A. Azzalini. *R package* `sn`*: The skew-normal and skew-t distributions (version 0.4-18)*. Università di Padova, Italia, 2013. URL `http://azzalini.stat.unipd.it/SN`.

A. Azzalini and A. Capitanio. Statistical Applications of The Multivariate Skew Normal Distribution. *Journal of The Royal Statistical Society Series B*, 61(3):579–602, 1999.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. Chapman & Hall, New York, 1984. ISBN 0-412-04841-8. URL http://www.crcpress.com/catalog/C4841.htm.

P. Bühlmann. Statistical Significance in High-Dimensional Linear Models. *Bernoulli*, 19 (4):1212–1242, 2013.

E. Cule and M. De Iorio. A Semi-Automatic Method to Guide the Choice of Ridge Parameter in Ridge Regression. *ArXiv e-prints*, May 2012.

B. Efron, R. Tibshirani, I. Johnstone, and T. Hastie. Least Angle Regression. *The Annals of Statistics*, 32(2):407–499, April 2004. ISSN 0090-5364. doi: 10.1214/009053604000000067. URL http://projecteuclid.org/Dienst/getRecord?id=euclid.aos/1083178935/.

J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

W. Fu and K. Knight. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28 (5):1356– 1378, 2000.

A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 1999.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

K. Shutes. *Non-Normality in Asset Pricing- Extensions and Applications of the Skew-Normal Distribution*. PhD thesis, University of Sheffield, 2004.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

L.-C. Wu, Z.-Z. Zhang, and D.-K. Xu. Variable Selection in Joint Location and Scale Models of the Skew-Normal Distribution. *Journal of Statistical Computation and Simulation*, pages 1–13, 2012. doi: 10.1080/00949655.2012.657198. URL http://www.tandfonline.com/doi/abs/10.1080/00949655.2012.657198.